A Central Limit Theorem for Parsimony Length of Trees

by

Mike Steel
Department of Mathematics and Statistics
University of Canterbury, Christchurch, New Zealand.
Larry Goldstein
and
Michael S. Waterman
Departments of Mathematics and
Molecular Biology
University of Southern California, Los Angeles, U.S.A.

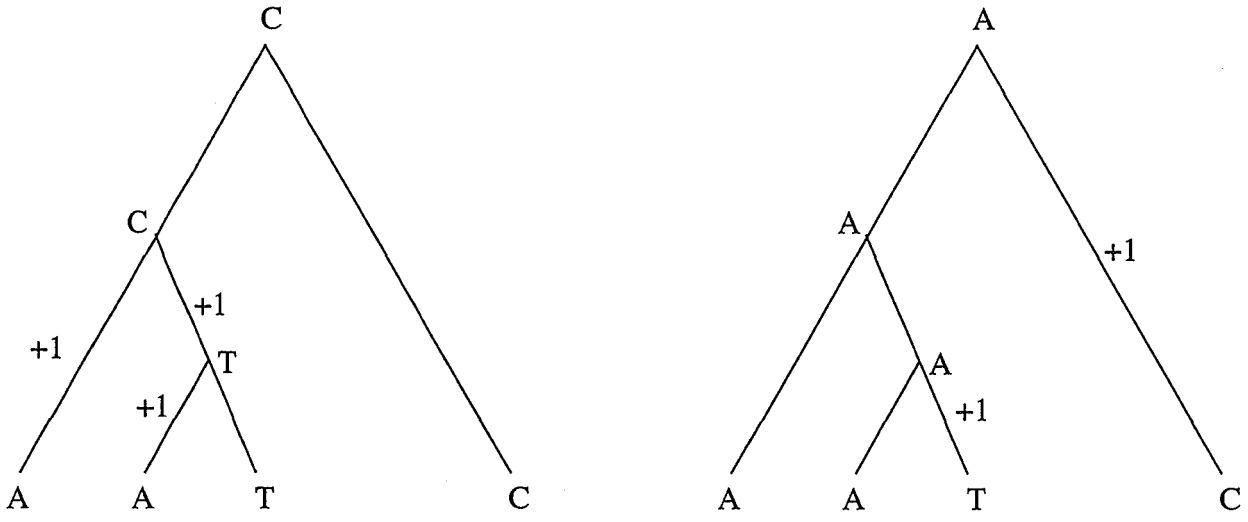
No. 115 October, 1994

Abstract : In phylogenetic analysis it is useful to study the distribution of parsimony length of a tree, under the null model by which the leaves are independently assigned letters according to prescribed probabilities. Except in one special case, this distribution is difficult to describe exactly. Here we analyse this distribution by providing a recursive and readily computable description, establishing large deviation bounds for the parsimony length of a fixed tree on a single site and for the minimum length (maximum parsimony) tree over several sites, and by showing that, under very general conditions, the former distribution converges asymptotically to the normal, thereby settling a recent conjecture. Furthermore, we show how the mean and variance of this distribution can be efficiently calculated. The proof of normality requires a number of new and recent results, as the parsimony length is not directly expressible as a sum of independent random variables, and so normality does not follow immediately from a standard central limit theorem.

1 Introduction

Parsimony is a commonly used method to provide a numerical measure of the fit of data to an evolutionary tree. Edwards and Cavalli-Sforza (1964) suggested the use of parsimony for evolutionary studies, and Fitch (1971) provided an algorithm to calculate parsimony scores for molecular sequences. Suppose that homologous sequences for n species have been aligned. For each position of the alignment, we consider a tree T with n leaves corresponding to the n species. Each leaf is given the letter that appears at the position in the sequence of the corresponding species. Letters are then placed at all interior vertices of the tree and the number of edges with different letters at the associated vertices is the score of that assignment. The parsimony score $L(T)$ is the minimum score for all possible assignments. The usual procedure is to find score $L_i(T)$ associated with position i and find the total score $\sum L_i(T)$ by summing over all positions. We will study the one position problem in most of this paper.

For a simple example we look at two assignments for a tree with $n = 4$ leaves.



The assignment on the left has score 3 while the assignment on the right has score 2 = $L(T)$. Optimal assignments are not always unique. To conform to standard terminology in graph theory, we will refer to the letters as *colors* and to an assignment as a *coloration* of the tree. We now reformulate the problem in more precise terms.

Throughout a *binary tree* will denote a tree $T = (V(T), E(T))$ which has labelled leaves of degree 1 and non-leaf vertices of degree 3. We will let n denote the number of leaves of T; hence $|E(T)| = 2n - 3$. A *rooted binary tree* is a binary tree with a subdivided edge, the resulting newly created vertex of degree two being the *root* of the tree (for technical reasons we also consider an isolated leaf as a rooted binary tree). Given a (possibly rooted) binary tree T, and a coloration $\mathcal{X}^*$ of $V(T)$ by a set of colors, the *changing number* of $\mathcal{X}^*$ on T is the number of edges of T whose ends are assigned different colors by $\mathcal{X}^*$. More especially, we will be interested in colorations of just the leaves of T. Given such a leaf coloration $\mathcal{X}$, the *length* of $\mathcal{X}$ on T, denoted $L(T)$, is the minimum changing number of any coloration $\mathcal{X}^*$ of $V(T)$ which extends $\mathcal{X}$. Such a coloration $\mathcal{X}^*$ is said to be a *minimal coloration* of T for $\mathcal{X}$. Note that a leaf coloration can have a large number of minimal colorations (indeed the number can grow exponentially with n - see Steel, 1993).

A particularly efficient, and for our purposes, useful way to calculate $L(T)$ is the forward version of Fitch's algorithm, which we now describe. If T is not already rooted, then define a root by choosing an arbitrary edge of T to subdivide. The value of $L(T)$ is not dependent on the choice of root. Direct all of the edges of T away from the root, so that each non-leaf vertex has two "children". Now, to each vertex T assign a pair (S, j), where S is a nonempty set of colors, and j is a non-negative integer, according to the following recursive scheme:

To leaf i, assign the pair $(\{\mathcal{X}(i)\}, 0)$.

To a vertex whose children have been assigned (S_1, j_1) and (S_2, j_2) assign the pair $(S_1 * S_2, j)$ where:

$$(S_1 * S_2, j) = \begin{cases} (S_1 \cap S_2, j_1 + j_2), & \text{if } S_1 \cap S_2 \neq \emptyset, \\ (S_1 \cup S_2, j_1 + j_2 + 1), & \text{otherwise.} \end{cases}$$

Eventually, pairs will be assigned to all of the vertices, including the root v whose associated pair we denote $(S(T), J)$. Hartigan (1973) established the following result:

Lemma 1 *$J = L(T)$ and $S(T) = \{\mathcal{X}^*(v) : \mathcal{X}^*$ is a minimal coloration of T for $\mathcal{X}\}$.*

After a parsimony score has been determined it must be evaluated. A natural procedure is to estimate the p-value of the score $L(T)$. That is, find the probability of observing a value as large or larger than $L(T)$ when the colors at the leaves are randomly assigned. In Section 2 we prove large deviations bounds of the form $\mathbb{P}[L(T) - \mathbb{E}[L(T)] > \lambda\sqrt{n}] \leq e^{-\lambda^2/2}$, that hold for **all** n for a fixed tree over a single site, and for the tree that minimizes the total parsimony score over several positions. In Section 3 we prove a central limit theorem for $L(T)$. Some special cases have been considered previously (Moon and Steel (1993)) but ours is the first general result. In particular, we allow the distribution of colors to vary from leaf to leaf, so that our results apply to sequences that exhibit variations in their base frequencies. In Section 4 we give recursions for computing the exact distribution of $L(T)$ in $O(n^2 4^c)$ steps, and for computing the mean and variance, $\mathbb{E}[L]$ and $V[L]$ in $O(n4^c)$ steps, where the alphabet size (number of colors) is c. In Section 5 we give an interesting example.

2 Large Deviation Bounds

We will consider first the single site model where each of the n leaves of the fixed tree T corresponds to the letter found in a given position in each of n (aligned) sequences. In this model, the leaves of a binary tree, T, are colored independently according to (possibly different) probability distributions. We will let π_i^α denote the probability that leaf i is assigned color α, π_i the probability distribution for leaf i, and $\boldsymbol{\pi} = \{\pi_i\}$ be the collection of the distributions for all the leaves.

In the special case where the leaves are bicolored according to identical and uniform distributions (*i.e.* $\pi_i^\alpha = 0.5$ for all i and both α), the distribution of $L(T)$ is dependent only on n but not on T and has been determined exactly by Steel (1993) and is given by

$$\mathbb{P}[L(T) = k] = \left[\binom{n-k}{k} + \binom{n-k-1}{k}\right] 2^{(k-n)}.$$

However, in general the exact distribution of $L(T)$ is complex, and the most one can hope for is either a recursive description (Section 3), an asymptotic expression as $n \to \infty$ (Theorem 2) or large

deviation bounds for finite n (Theorem 1). Regarding large deviation bounds we have the following results, in which $\mathbb{E}[L] = \mathbb{E}[L(T)]$ and $V[L] = Var[L(T)]$.

Theorem 1 *Given T and $\boldsymbol{\pi}$,*

$$\mathbb{P}[L(T) - \mathbb{E}[L] > \lambda\sqrt{n}] \leq e^{-\lambda^2/2}, \tag{1}$$

$$\mathbb{P}[L(T) - \mathbb{E}[L] < -\lambda\sqrt{n}] \leq e^{-\lambda^2/2}, \tag{2}$$

and for each $p > 0$

$$\mathbb{E}\left[\left|\frac{L(T) - \mathbb{E}[L]}{\sqrt{n}}\right|^p\right] \leq 2p\int_0^\infty \lambda^{p-1}e^{-\lambda^2/2}d\lambda. \tag{3}$$

In case $p = 2$, this can be improved to

$$V[L]/n \leq 1/2. \tag{4}$$

Proof. First we verify that the parsimony length L satisfies the Lipschitz condition

$$|L(X_1 \cdots X_{i-1}, X_i, X_{i+1}, \ldots, X_n) - L(X_1, \ldots, X_{i-1}, Y_i, X_{i+1}, \ldots, X_n)| \leq 1$$

for all Y_i, where here $L(X_1, \ldots, X_n)$ is the parsimony length of T when leaf i is assigned color X_i, $i = 1, \ldots, n$. It suffices by symmetry to establish

$$L(X_1, \cdots, X_{i-1}, Y_i, X_{i+1}, \ldots, X_n) \leq L(X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n) + 1 \tag{5}$$

The length $L(X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n)$ is the changing number of a particular coloration of all the vertices of T. Altering the color of leaf i from X_i to Y_i increases this number by at most 1, and yields a coloration that extends $(X_1, \ldots, X_{i-1}, Y_i, X_{i+1}, \ldots, X_n)$. Equation (5) follows by the minimality of $L(X_1, \ldots, X_{i-1}, Y_i, X_{i+1}, \ldots, X_n)$. Now, using the independence of $X_1, X_2, \ldots, X_n$,

equations (1) and (2) follow by applying the Azuma-Hoeffding inequality as it appears in Theorem 4.2 pg. 90 of Alon and Spencer. Equations (1) and (2) imply $\mathbb{P}[|L - \mathbb{E}[L]| > \lambda\sqrt{n}] \leq 2e^{-\lambda^2/2}$, and equation (3) now follows from

$$\mathbb{E}[W^p] = p\int_0^\infty \lambda^{p-1} P(W > \lambda) d\lambda$$

for any $W \geq 0$.

The bound for $V[L(T)]$ follows immediately from Steele (1986), applied to the Lipschitz condition described above. ■

Again consider the case where there are n aligned sequences, each now of length $k \geq 1$, where all sites are independent and generated according to our single site model. Let $X_{ij}, i = 1, \ldots, n, j = 1, \ldots, k$, be the color assigned to sequence i at position (site) j, and let $L^* = L^*([X_{ij}])$ be the length of a minimum length (maximum parsimony) tree for this data, *i.e.*

$$L^* = \min_T \sum_{j=1}^{k} L(T, \mathcal{X}_j)$$

where $\mathcal{X}_j(i) = X_{ij}$. We then have the following

Theorem 2 *Under the independent multisite model described above,*

$$\mathbb{P}[L^* - \mathbb{E}[L^*] > \lambda\sqrt{nk}] \leq e^{-\lambda^2/2}, \tag{6}$$

$$\mathbb{P}[L^* - \mathbb{E}[L^*] < -\lambda\sqrt{nk}] \leq e^{-\lambda^2/2}. \tag{7}$$

Proof. Suppose $X'_{ij} = X_{ij}$ except for one value (i_0, j_0) of (i, j). As in the proof of Theorem 1, since the nk X_{ij}'s are independent, we need only verify

$$|L^* - L'| \leq 1 \tag{8}$$

where $L' = L^*([X'_{ij}])$. By symmetry it suffices to show

$$L' \leq L^* + 1. \tag{9}$$

Suppose T is a minimum length (maximum parsimony) tree for $[X_{ij}]$. Then the length of T for $[X'_{ij}]$ is:

$$\sum_{\substack{j=1,\ldots,k \\ j\neq j_0}} L(T,\mathcal{X}_j) + L(T,\mathcal{X}'_{j_0})$$

where $\mathcal{X}'_{j_0}(i) = X_{ij_0}$.

Now, $\mathcal{X}'_{j_0}$ differs from $\mathcal{X}_j$ in exactly one coordinate, and so, from equation (5) we have that

$$L(T,\mathcal{X}'_{j_0}) \leq L(T,\mathcal{X}_{j_0}) + 1.$$

Thus, the length of T for $[X'_{ij}]$ is at most

$$\sum_{j=1}^{k} L(T,\mathcal{X}_j) + 1 = L^* + 1$$

and now (9) follows by the minimality of T. ■

3 Central Limit Theorem

We turn now to the asymptotic behavior of $L(T)$. In the special 2-color case described earlier ($\pi_i^\alpha = 0.5$ for all i and both α), $L(T)$ was shown to be asymptotically normal (Moon and Steel, 1993). In general however, if no restrictions are placed on the distributions $\boldsymbol{\pi} = \{\pi_i^\alpha\}$ then $L(T)$ need not be asymptotically normal, since the distribution is obviously degenerate if $\pi_i^\alpha \in \{0,1\}$. Thus, in order to explore asymptotically the distribution of $L(T)$ we bound the π_i^α's uniformly away from 0; that is we assume:

$$\pi_i^\alpha > \epsilon, \quad \text{for all } i,\alpha. \tag{10}$$

for some $\epsilon > 0$ (independent of n).

A conjecture, which generalizes conjectures reported by Archie and Felsenstein (1993) and Moon and Steel (1993), is that $L(T)$ should be asymptotically normal under condition (10). The following theorem, proved in section 2, shows that this is indeed so, and provides order estimates for the

growth in the mean and variance of the distribution. Note that, under assumption (10), a quantity closely related to $L(T)$, namely the root set $S(T)$, can still be degenerate asymptotically, as the example in Section 4 (below) shows. In the following theorem recall that n denotes the number of leaves of T.

Theorem 3 *Assuming (10), as $n \to \infty$, the distribution of $\frac{L(T)-\mathbb{E}[L]}{\sqrt{V[L]}}$ converges to the standard normal distribution $N(0,1)$. Furthermore, both $\mathbb{E}[L]$ and $V[L]$ grow (approximately) linearly with n.*

For completeness, and due to its central role, we now make precise as a separate proposition the last part of the claim in Theorem 3.

Proposition 1 *For all π satisfying (10), there exists a $\delta > 0$ (dependent on ϵ) such that $[\delta, 1)$ contains $\frac{\mathbb{E}[L]}{n}$ and $\frac{V[L]}{n}$ for all binary trees T.*

In order to prove this proposition and the theorem, we need to establish a number of preliminary results; the first three of which are purely combinatorial properties of binary trees.

Lemma 2 ("Lonely leaves lemma") *The leaves of any binary tree T can be ordered $l_1, l_2, \ldots, l_n$ in such a way that, for at least $1 + n/3$ values of i, l_i and l_{i-1} are separated by no more than 3 edges.*

Proof. For leaves i and j let $d(i,j)$ denote the number of edges of T separating i and j. For $n > 3$, delete from T all its leaves, and their incident edges, to obtain a tree T_1 (which is the subdivision of a unique binary tree T_2), as in Figure 1. Note that an edge e of T_2 corresponds to a path in T_1 and we denote by $X(e)$ the (possibly empty) set of leaves of T which are adjacent to any vertex in this path. We partition the edges of T_2 into four classes as follows:

C_1: Edges incident with a leaf of T_2

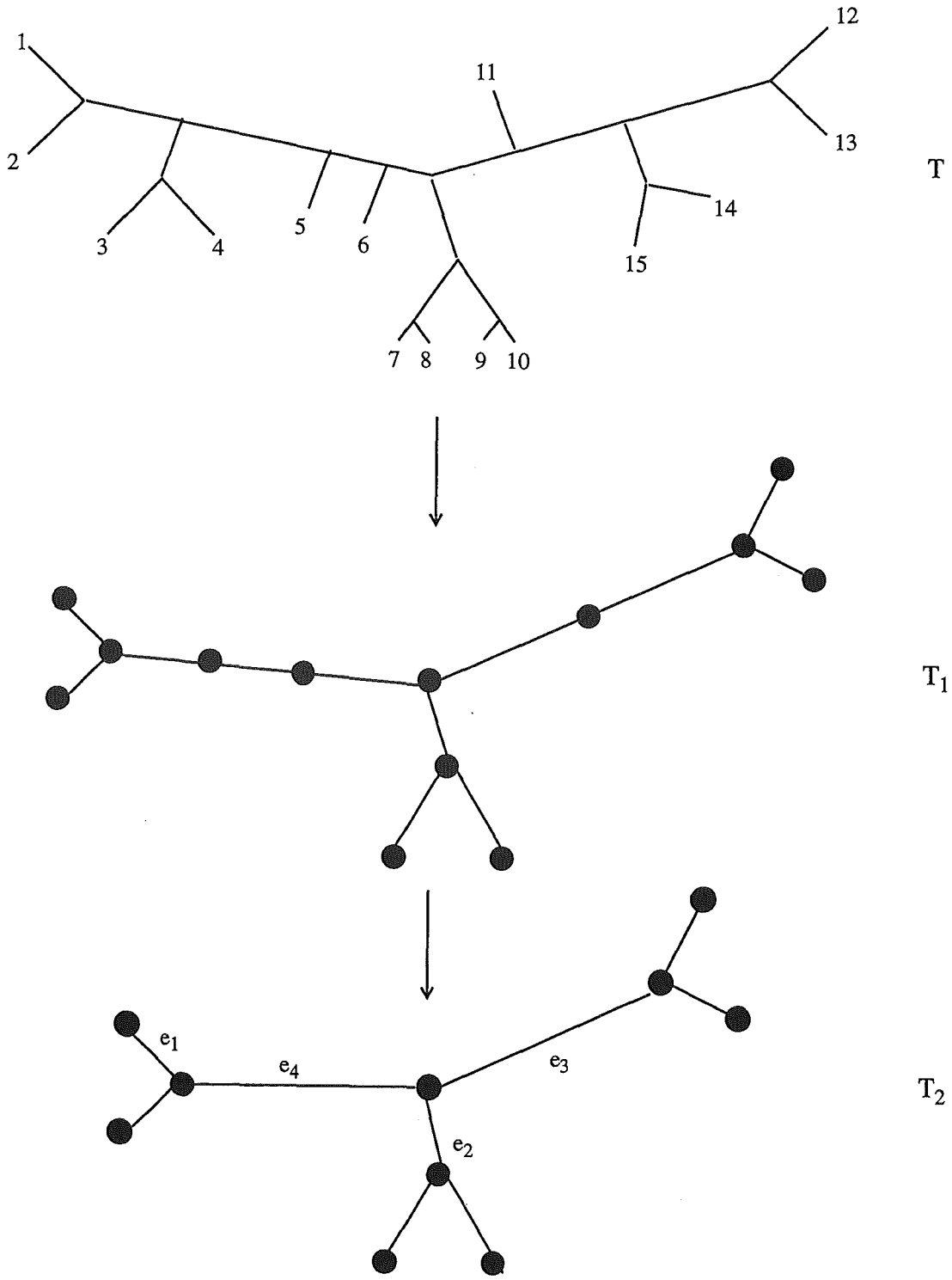


Figure 1: A derived tree T_2 with examples e_i of edges in class C_i.

C_2: Edges not incident with a leaf of T_2 and with $|X(e)| = 0$

C_3: Edges not incident with a leaf of T_2 and with $|X(e)| = 1$

C_4: Edges not incident with a leaf of T_2 and with $|X(e)| \geq 2$

For example the tree in Figure 1 has $C_2 = \{e_2\}$, $C_3 = \{e_3\}$, $C_4 = \{e_4\}$. Note the sets $X(e)$ from cases (1), (3) and (4) partition the leaves of T, and case (3) covers precisely the "lonely" leaves i for which $d(i,j) > 3$ for all leaves j. It is clear that we can relabel the leaves of T as $l_1, \ldots, l_n$ in such a way that for any edge e in class (1) or (4), $d(l_i, l_{i-1}) \leq 3$ for at least $|X(e)| - 1$ values of i. Under this ordering the number of leaves i for which $d(l_i, l_{i-1}) \leq 3$ is precisely:

$$\sum_{e \in C_1 \cup C_4} (|X(e)| - 1) \geq \sum_{e \in C_1 \cup C_4} |X(e)|/2 \geq (n - |C_2| - |C_3|)/2 \geq (n - |C_3|)/2.$$

Now, $|C_3|$ is bounded above by the number of edges of T_2 that are not incident with a leaf. Thus, if T_2 has k leaves, then $|C_3| \leq k - 3$, and futhermore, the number n of leaves of T, is at least $2k + |C_3|$, thereby providing the inequality: $n \geq 2(|C_3| + 3) + |C_3|$. Thus, $|C_3| \leq n/3 - 2$, and so, $(n - |C_3|)/2 \geq n/3 + 1$. Combining this with the string of inequalities above establishes the Lemma, and actually shows, moreover, that this bound is best possible. ■

Definition: Suppose T is a binary tree leaf-labelled by L. A *leaf-covering forest for* T is a collection of vertex disjoint subtrees of T whose leaf sets form a partition of the leaf set of T.

Lemma 3 ("Tree-chopping lemma") *For any binary tree T, and integer $k > 0$, T has a leaf covering forest F, for which the number of leaves in each tree in F is at most $2k - 2$, and, except possibly for one tree, is also at least k.*

Proof. Select a leaf i of T, and let T' denote a subtree of T containing i. Direct all edges of T' away from i, thereby creating a partial order $\leq$ on the vertices of T' with minimal element i. Thus, each vertex $v \in V[T']$ has a set of "descendents" $= \{v' \in V[T'] : v \leq v'\}$, and we let $d(T', v)$ denote the number of **leaves** of T' which are descendents of v. The following algorithm gives the required leaf covering forest F.

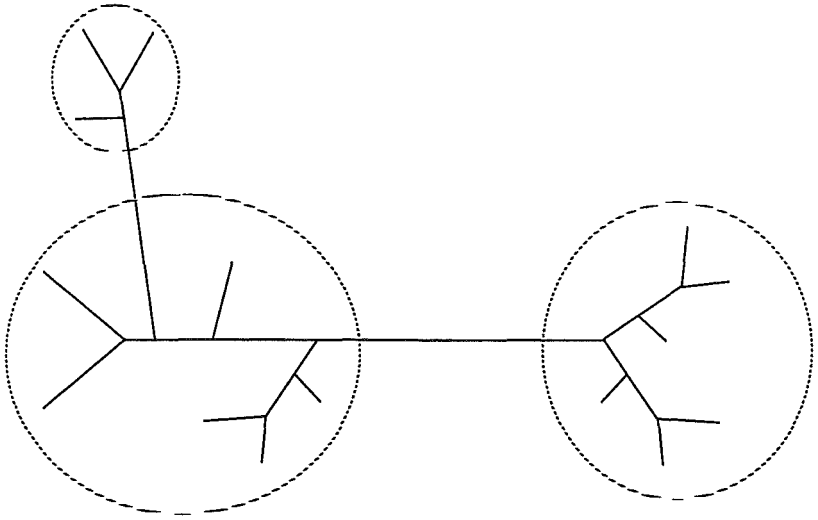


Figure 2: A leaf-covering forest (circled) for $k = 4$.

If T has less than k leaves the lemma holds. Otherwise there is a maximal (under $\leq$) vertex v of T with $d(T, v) \geq k$. It also is true that $d(T, v) \leq 2k - 2$. Otherwise $d(T, v) > 2k - 2$ and, if u, u' are the descendents of v,

$$d(T, v) = d(T, u) + d(T, u').$$

Then $\max\{d(T, u), d(T, u')\} \geq k$, contradicting the maximality of v. Next we remove the tree consisting of v, and its descendent vertices and incident edges and place it in F. Then we inductively repeat the above procedure to the remaining tree until it has less then k leaves. Note that the subtrees removed can have vertices of degree 2. (See Figure 2). ■

Figure 2 shows a leaf-covering forest of T in case $k = 4$.

Lemma 4 *Suppose $F = \{T_1, \ldots, T_r\}$ is a leaf-covering forest for T. Given a leaf coloration $\mathcal{X}$ of T, let $\mathcal{X}_i$ denote the restriction of $\mathcal{X}$ to the leaves of T that lie in T_i, and let*

$$\Delta = L(T) - \sum_{i=1}^{r} l(\mathcal{X}_i, T_i).$$

Then $0 \leq \Delta \leq r - 1$.

Proof of Lemma 4. First we note that the forest has no more edges joining the T_i's than the number of edges of a binary tree with n leaves. Therefore $\Delta \leq 2r-3$ which is sufficient for proving Theorem 2. The better bound of the lemma might be useful in other contexts. We first establish the following sublemma: Suppose T is any tree (possibly with degree 2 vertices) and F is a subforest of T, whose components collectively cover all T's leaves. Then F determines a collection Q of subtrees of T, as follows: Let E' denote the set of edges of T that do not lie entirely in F, and let V' denote the vertices of T that are incident with an edge of E'. Let Q = Q(T,F) denote the set of (leaf-overlapping) subtrees of (V', E') that have all their leaves, but no other vertices consisting of vertices from trees in F. An example of this construction is given in Figure 3, where the five trees in F are circled. Let n_t denote the number of leaves of $t \in Q$. We claim that:

$$\sum_{t \in Q}(n_t - 1) = |F| - 1.$$

To establish ths claim, we first note that we may, without loss of generality, assume that F consists entirely of isolated vertices.

Then, take any vertex v in F and direct all edges of T away from v. Then each vertex v' in $F - \{v\}$ has an incident edge $e_{v'}$ directed towards it, and $e_{v'}$ lies in precisely one of the trees $t_{v'}$ in Q, and v' is a leaf of $t_{v'}$. All but one of the leaves of $t_{v'}$ is identified in this way, thus we have a one-to-one mapping from $F - \{v\}$ to the one-vertex-deleted leaf sets of the trees in Q, and this establishes the claim.

Now let $\mathcal{X}^*$ be a minimal coloration of T for $\mathcal{X}$. Let $\mathcal{X}_i^*$ denote the restriction of $\mathcal{X}^*$ to $V(T_i)$. Then $\mathcal{X}_i^*$ extends $\mathcal{X}_i$ and so $l(\mathcal{X}_i, T_i)$ is at most the changing number of $\mathcal{X}_i^*$ on T_i, hence

$$\sum_{i=1}^{r} l(\mathcal{X}_i, T_i) \leq \text{ changing number of } \mathcal{X} \text{ on } T = l(\mathcal{X}, T)$$

which shows that $0 \leq \Delta$. To obtain an upper bound for Δ, let $\mathcal{X}^i$ denote a minimal coloration of T_i for $\mathcal{X}_i$. Define a coloration $\mathcal{X}'$ of the vertices of T that lie in trees from F, by setting: $\mathcal{X}'(v) = \mathcal{X}^i(v)$ if $v \in V(T_i)$. Extend $\mathcal{X}'$ to a coloration $\mathcal{X}''$ of $V(T)$ by coloring any vertices in T not covered by

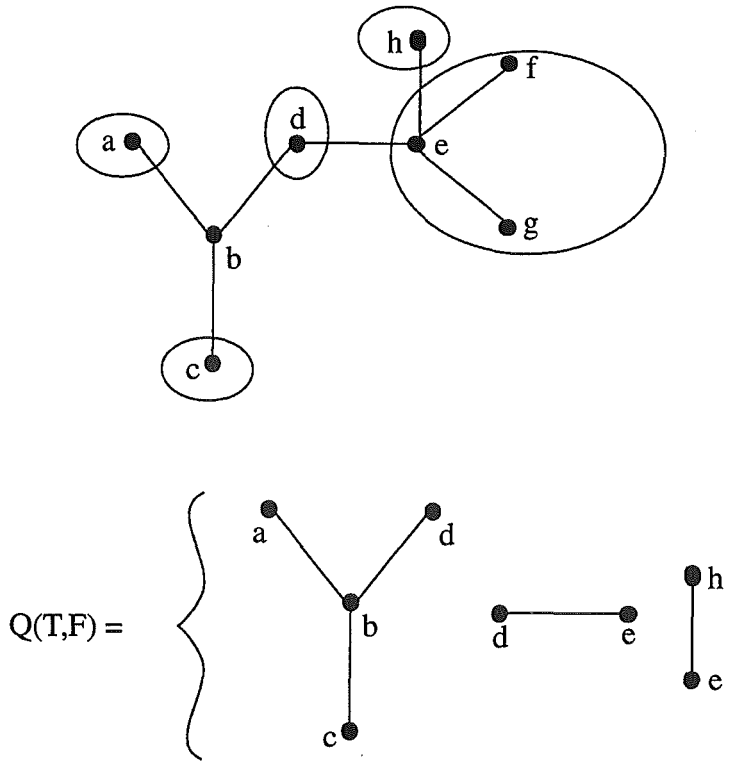


Figure 3: Subtrees obtained from a leaf covering forest of five trees (circled)

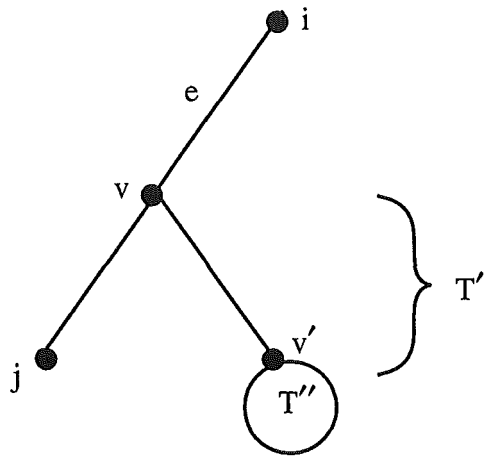


Figure 4: Tree decomposition

trees in F (and therefore lying in $Q(T,F)$) in such a way that all the non-leaf vertices in any component of $Q(T,F)$ are assigned the same color as one of the leaves of that component. The changing number of $\mathcal{X}''$ on T is the sum (over i) of the changing numbers of $\mathcal{X}^i$ on T_i, plus the sum of the changing numbers of the restriction of $\mathcal{X}''$ to the components of $Q(T,F)$. This latter sum is no more than $\sum_{t\in Q}(n_t-1)$. But this sum was shown above to equal $r-1$, so that $\mathcal{X}''$ has changing number at most $\sum_{i=1}^{r} l(\mathcal{X}_i,T_i)+r-1$, and this gives the required upper bound on $l(\mathcal{X},T)$ since $\mathcal{X}''$ extends $\mathcal{X}$. ∎

Proof of Proposition 1. The upper bounds are very easily derived, however the justification for the lower bounds, particularly for V[L] is much more involved and occupies most of the proof. For any leaf coloration $\mathcal{X}$, consider the coloration $\mathcal{X}^*$ of $V(T)$ which assigns the most frequently occurring leaf color to all the vertices of T. Then $\mathcal{X}^*$ extends $\mathcal{X}$, and has changing number $n-a$, where a is the number of leaves colored with the most frequently occurring color. Thus $L \leq \left(1-\frac{1}{c}\right)n$ and

$$\frac{\mathbb{E}[L]}{n} \leq 1-\frac{1}{c} < 1,$$

where c is the number of colors.

To obtain a lower bound, consider any leaf i of T, and the edge e incident with i. Deleting e and i from T produces a rooted binary tree T' possessing a vertex v of degree 2, that was formerly incident with e. Let $\mathcal{X}'$ denote the restriction of $\mathcal{X}$ to T'. Applying Lemma 1, we see that,

$$L(T) = L(T') + D_i$$

where D_i is a 0,1 random variable, which equals 1 precisely if $\mathcal{X}(i) \notin S(T')$. Taking expectation,

$$\mathbb{E}[L(T)] = \mathbb{E}[L(T')] + \mathbb{P}[D_i = 1].$$

Now a modification of the Example in Section 4 shows that, in general, $\mathbb{P}[D_i = 1]$ can be arbitrarily close to 0. However we can always find, in any binary tree, a leaf i which is separated from another leaf j by just two edges, and in this case we will show that $\mathbb{P}[D_i = 1]$ is bounded away from 0. Thus, represent T and T' as in Figure 4, and let T'' be the rooted tree obtained from T' by deleting leaf j, the root and its two incident edges. See Figure 4. From Lemma 1, we have, for any $\alpha \neq \beta$:

$$\mathbb{P}[D_i = 1] = \mathbb{P}[\mathcal{X}(i) \notin S(T')] \geq \mathbb{P}[\mathcal{X}(i) = \alpha \& S(T') = \{\beta\}]$$

and by the independence condition (1),

$$\begin{aligned} \mathbb{P}[\mathcal{X}(i) = \alpha \& S(T') = \{\beta\}] &= \mathbb{P}[\mathcal{X}(i) = \alpha] \times \mathbb{P}[S(T') = \{\beta\}] \\ &\geq \epsilon \mathbb{P}[S(T') = \{\beta\}]. \end{aligned}$$

Combining these two inequalities we have, for any β:

$$\mathbb{P}[D_i = 1] \geq \epsilon \mathbb{P}[S(T') = \{\beta\}]. \tag{11}$$

Now,

$$\begin{aligned} \mathbb{P}[S(T') = \{\beta\}] &\geq \mathbb{P}[\mathcal{X}(j) = \beta \& \beta \in S(T'')] = \\ &\mathbb{P}[\mathcal{X}(j) = \beta] \times \mathbb{P}[\beta \in S(T'')] \geq \epsilon \mathbb{P}[\beta \in S(T'')]. \end{aligned} \tag{12}$$

Now, since D_k is a 0,1 random variable,

$$\mathbb{E}[D_k|\mathcal{F}_k] = \mathbb{P}[D_k = 1|\mathcal{F}_k] = \sum_{\substack{\alpha,X:\\ \alpha\notin X}} \mathbb{I}\{X_k = \alpha\}\mathbb{P}[S(T') = X|\mathcal{F}_{k-1}]. \tag{14}$$

Now,

$$\begin{aligned}
\mathbb{E}[d_k^2] &\geq \mathbb{E}[d_k^2|X_k = \alpha \& Y_k = \alpha' \& X_{k-1} = \alpha] \times \mathbb{P}[X_k = \alpha, \& Y_k = \alpha' \& X_{k-1} = \alpha] \\
&\geq \epsilon^3 \mathbb{E}[d_k^2|X_k = \alpha \& Y_k = \alpha' \& X_{k-1} = \alpha] \\
&\geq \epsilon^3 \mathbb{E}[d_k|X_k = \alpha \& Y_k = \alpha' \& X_{k-1} = \alpha]^2 = \epsilon^3 E_{\alpha,\alpha'}^2, \text{ say.}
\end{aligned}$$

Consider the contribution to $E_{\alpha,\alpha'}$ given by the first term of d_k in equation (13):

$$\mathbb{E}\Big[\mathbb{E}[D_k|\mathcal{F}_k]\Big|X_k = \alpha, Y_k = \alpha', X_{k-1} = \alpha\Big]$$

Using equation (14) this equals

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\substack{\alpha,X\\ \alpha\notin X}} \mathbb{I}(X_k = \alpha)\mathbb{P}[S(T') = X|\mathcal{F}_{k-1}] \;\Big|\; X_k = \alpha, Y_k = \alpha', X_{k-1} = \alpha\right] \\
= \;&\mathbb{E}\left[\sum_{\substack{X\\ \alpha\notin X}} \mathbb{P}[S(T') = X|\mathcal{F}_{k-1}] \;\Big|\; X_{k-1} = \alpha\right] \\
= \;&\mathbb{E}\left[\mathbb{P}[\alpha \notin S(T')|\mathcal{F}_{k-1}]|X_{k-1} = \alpha\right] \\
= \;&\mathbb{P}[\alpha \notin S(T')|X_{k-1} = \alpha].
\end{aligned}$$

An analogous argument applies to the contribution to $E_{\alpha,\alpha'}$ given by the second term in (13), to show

$$\begin{aligned}
E_{\alpha,\alpha'} &= \mathbb{P}[\alpha \notin S(T')|X_{k-1} = \alpha] - \mathbb{P}[\alpha' \notin S(T')|X_{k-1} = \alpha], \\
&= \mathbb{P}[\alpha' \in S(T')|X_{k-1} = \alpha] - \mathbb{P}[\alpha \in S(T')|X_{k-1} = \alpha].
\end{aligned}$$

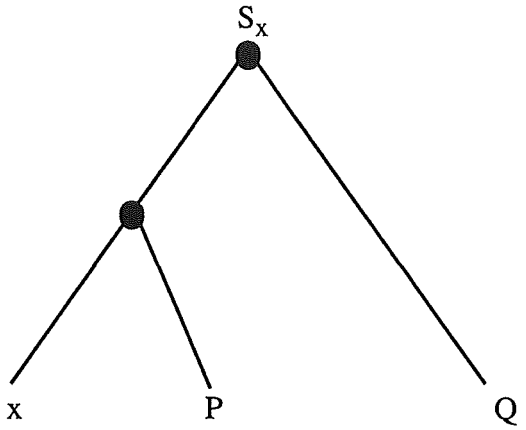


Figure 5:

Now if l_k and l_{k-1} are separated by either 2 or 3 edges Lemma 5, proved below, shows that there exist α and α' so that $|E_{\alpha,\alpha'}| > c$ for some constant $c > 0$ (independent of n) no matter what distribution we have over the remaining leaves. Thus, $\mathbb{E}[d_k^2] \geq c^2\epsilon^3$. Since there are at least $n/3+1$ such values of k for which this is so, (by Lemma 2) it follows that $V[L] \geq c^2 n\epsilon^3/3$, which provides the required lower bound. This completes the proof of the proposition. ■

Lemma 5 *If leaves l_k and l_{k-1} are separated by either 2 or 3 edges, then α and α' can be chosen so that $|E_{\alpha,\alpha'}| > c$ for some constant $c > 0$ independent of n.*

Proof of Lemma 5. Let

$$S_x = (\{x\} * P) * Q$$

where P, Q are random variables taking values in $2^{\mathcal{C}} - \phi$ (the nonempty subsets of the set of colors $\mathcal{C}$) and $x \in \mathcal{C}$. If l_k and l_{k-1} are separated by three edges, then taking P, Q to be the root sets of the two rooted subtrees on the path connecting l_{k-1} and l_k we have that

$$E_{\alpha,\beta} = \mathbb{P}[\beta \in S_\alpha] - \mathbb{P}[\alpha \in S_\alpha].$$

If l_k and l_{k-1} are separated by just two edges then this same equation applies if we take P to be the root set of the rooted subtree between l_{k-1} and l_k, and if we take Q to be $\mathcal{C}$ with probability 1.

Thus to establish the Lemma it suffices to show that there exists $\varepsilon = \varepsilon(|\mathcal{C}|) > 0$, such that for any distribution on P, Q, there exists a pair $\alpha, \beta \in \mathcal{C}$ with

$$|E_{\alpha,\beta}| > \varepsilon.$$

Let $x_p = \mathbb{P}[P = p], p \in 2^{\mathcal{C}} - \phi$; $y_q = \mathbb{P}[Q = q], q \in 2^{\mathcal{C}} - \phi$, and let $\phi := \sum_{\substack{\alpha,\beta \\ \alpha \neq \beta}} E^2_{\alpha,\beta} \geq 0$. Note that ϕ is a continuous function of the $\{x_p\}$ and $\{y_p\}$ (since $\phi = \sum_{\alpha,\beta}(\sum_{p,q} \lambda_{pq} x_p y_q)^2$ for suitable coefficients $\lambda_{pq} = 0, \pm 1$) and that $\{x_p, y_q\}$ are constrained to lie in the **closed** set C:

$$C = \begin{cases} x_p \geq 0, \ y_p \geq 0 \\ \sum_p x_p = \sum y_p = 1. \end{cases}$$

By continuity of ϕ and closure of C, to establish the claim it suffices to show that ϕ is never 0 within C - that is, we can, for any distribution on P, Q, find a pair α, β such that

$$E_{\alpha,\beta} \neq 0.$$

We suppose not to derive a contradiction. Thus, suppose for some distribution that $E_{\alpha,\beta} = 0$ for all α, β. In particular

$$\begin{aligned} & E_{\alpha,\beta} + E_{\beta,\alpha} = 0 \\ \Rightarrow \ & \underbrace{\mathbb{P}[\alpha \in S_\beta] - \mathbb{P}[\alpha \in S_\alpha]} + \underbrace{\mathbb{P}[\beta \in S_\alpha] - \mathbb{P}[\beta \in S_\beta]} = 0 \\ \Rightarrow \ & \Delta_{\alpha,\beta} + \Delta_{\beta,\alpha} = 0 \end{aligned}$$

Now, for any choice $P = p, Q = q$ we have

$$\alpha \in S_\beta \Rightarrow \alpha \in S_\alpha \tag{15}$$

(since if $\alpha \notin S_\alpha$ we must have $P \cap Q = \emptyset$ and $\alpha \notin P \cup Q$, and $\alpha \notin P \cup Q \Rightarrow \alpha \notin S_\beta$). Consequently $\Delta_{\alpha,\beta} \leq 0$, and hence $\Delta_{\beta,\alpha} \leq 0$ by symmetry. Thus $\Delta_{\alpha,\beta} = 0$ and in order to obtain a contradiction it suffices to find α, β such that

$$\Delta_{\alpha,\beta} < 0.$$

We distinguish two cases:

(I) There exists $\alpha : \mathbb{P}[\alpha \in P] > 0$ and $\mathbb{P}[\alpha \notin Q] > 0$.

(II) For all $\alpha : \mathbb{P}[\alpha \in P] = 0$ or $\mathbb{P}[\alpha \notin Q] = 0$.

In case (I) select β so that $\mathbb{P}[\alpha \notin Q, \beta \in Q] > 0$. Let E be the event $\alpha \in P, \alpha \notin Q, \beta \in Q$. Then $\mathbb{P}[E] > 0$ and

$$\begin{aligned} \mathbb{P}[\alpha \in S_\alpha | E] &= 1 \\ \mathbb{P}[\alpha \in S_\beta | E] &= 0. \end{aligned}$$

Since $\mathbb{P}[\alpha \in S_\alpha | P = p, Q = q] \geq \mathbb{P}[\alpha \in S_\beta | P = p, Q = q]$ for all p, q, by implication (15), we deduce that

$$\Delta_{\alpha,\beta} < 0,$$

the required contradiction.

For case (II) we consider the possible subcases:

(i) There exist $\alpha, \beta : \mathbb{P}[\alpha \in P] = 0,\ \mathbb{P}[\beta \notin Q] = 0$

(ii) For all $\alpha : \mathbb{P}[\alpha \notin Q] = 0$

(iii) For all $\alpha : \mathbb{P}[\alpha \in P] = 0$ (impossible!)

In case (i) $\mathbb{P}[\alpha \in S_\beta] = 0, \mathbb{P}[\beta \in S_\beta] = 1$ so $E_{\beta,\alpha} \neq 0$, the required contradiction.

In case (ii) we have $\mathbb{P}[Q = C] = 1$ so select any $\alpha \in P$ for which $\mathbb{P}[\alpha \in P] > 0$ to obtain $E_{\alpha,\beta} \neq 0$ for any $\beta \neq \alpha$. Case (iii) cannot arise. This completes the proof. ∎

Proof of Theorem 2. An outline of the proof is as follows: We use the tree-chopping lemma to construct a family of comparably-sized disjoint subtrees of T, the sum of whose intrinsic parsimony

lengths approximates $L(T)$ via Lemma 4. It is important that in chopping up T the component subtrees grow in size sufficiently quickly, but not as fast as the number of them. In this way, we can apply a version of the central limit theorem, due to Liapunov, for double arrays of random variables, and verify its hypotheses using the Proposition and Theorem 1.

The required central limit theorem (see Serfling (1980)) states the following: For each n, let $X_{n1}, \ldots, X_{nk}$ be $k = k(n)$ independent random variables with finite p-th moments for some $p > 2$. Let

$$A_n = \sum_j \mathbb{E}[X_{nj}]; B_n = \sum_j Var[X_{nj}]. \tag{16}$$

If

$$B_n^{-p/2} \sum_j \mathbb{E}[|X_{nj} - \mathbb{E}[X_{nj}]|^p] \to 0 \text{ as } n \to \infty, \tag{17}$$

then $W_n = \frac{\sum_j X_{nj} - A_n}{\sqrt{B_n}}$ converges to the standard normal distribution, written $W_n \to N(0,1)$, as $n \to \infty$.

We apply this theorem as follows. Firstly, use Lemma 3 with $k(n) = \lfloor n^\alpha \rfloor$ where $\alpha > 0.5$ to construct a leaf-covering forest $F = \{T_1, \ldots, T_r\}$ for T which satisfies the constraints prescribed by Lemma 3 for $k = k(n)$. The number r of trees in F is $O(n^\beta)$, $\beta = 1 - \alpha$, meaning that r/n^β is contained in a fixed positive interval for all trees. Let $X_{nj} = L(T_j)$, as in Lemma 4. Consider the two quantities:

$$\begin{aligned} Z_T &= \frac{L(T) - \mathbb{E}[L]}{\sqrt{V[L]}}; \\ W_T &= \frac{\sum_j X_{nj} - A_n}{\sqrt{B_n}} \end{aligned}$$

where A_n and B_n are given by equation (16). Then T_j has at most $2k(n) - 2$ leaves, and (3) of Theorem 1 shows that $\mathbb{E}[|X_{nj} - \mathbb{E}[X_{nj}]|^p] = O(n^{\alpha p/2})$. Also by the lower bound on the variance

given by the Proposition, we have $B_n \geq r\delta k(n) > c'n$ for some constant $c' > 0$ independent of n. Thus,

$$B_n^{-p/2} \sum_j \mathbb{E}[|X_{nj} - \mathbb{E}[X_{nj}]|^p] = O((c'n)^{-p/2} \times n^{\beta} n^{\alpha p/2}) = O(n^{\beta(1-p/2)})$$

which coverges to 0 as $n \to \infty$ when $p > 2$. Thus, condition (17) is satisfied. Furthermore, for each n, $X_{n1}, \ldots, X_{nr}$ are independent. Thus, the central limit theorem described above applies, to show that $W_T \to N(0,1)$. Now,

$$Z_T = \sqrt{\frac{B_n}{V[L]}} W_T + \frac{\Delta - \mathbb{E}[\Delta]}{\sqrt{V[L]}} \tag{18}$$

where Δ is defined by Lemma 4. Again, invoking the lower bound for the variance (given by the Proposition), this time for $V[L]$, and the upper bound from Lemma 4, $|\Delta| < r = O(n^{\beta})$, we see that the second term in (18) converges in probability to 0. Regarding the first term, note that, from the definition of Δ we have:

$$V[L] = B_n + Var[\Delta] + 2Cov\left[\Delta, \sum_j X_{nj}\right].$$

Applying the Cauchy-Schwartz inequality, and again using the bound, $|\Delta| < r = O(n^{\beta})$ we have:

$$V[L] = B_n + O(n^{2\beta}) + O(n^{\beta}\sqrt{B_n})$$

so that $B_n/V[L]$ converges to 1 as $n \to \infty$ since $B_n > c'n$. Thus, $\sqrt{\frac{B_n}{V[L]}}$ converges to 1 in probability, and so, since $W_T \to N(0,1)$, we can apply Slutsky's lemma (Durrett (1991)) and deduce that $Z_T \to N(0,1)$, as required. ∎

4 The Exact Distribution and Its Mean and Variance

Application of Theorem 1 depends on knowledge of $\mathbb{E}[L]$ and $V[L]$, the mean and variance of $L = L(T)$. In this section we present efficient recursions for calculating these quantities, given T and its leaf distribution $\boldsymbol{\pi}$. First however we give an algorithm that is polynomial in n for computing the exact distribution of $L = L(T)$.

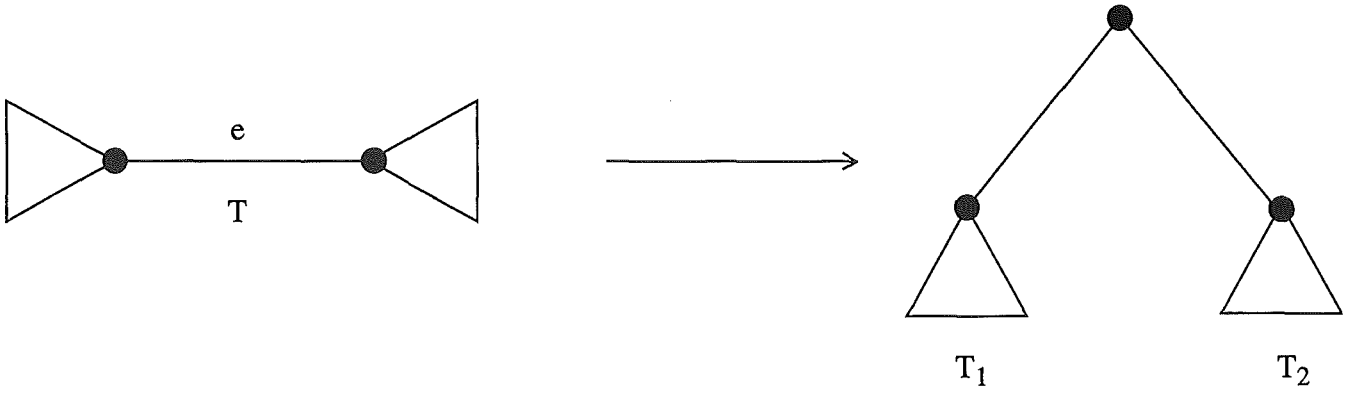
Let $\mathbb{P}[T, \pi, k]$ denote the probability that $L(T) = k$. To obtain a recursive formula, we need more generally to consider, for each nonempty subset X of colors, the quantity

$$\mathbb{P}_X[T, \pi, k] = \mathbb{P}[L(T) = k \text{ and } S(T) = X].$$

Now consider the ordinary generating function:

$$F_X(T, \pi, x) = \sum_{k \geq 0} \mathbb{P}_X[T, \pi, k] x^k.$$

Subdivide an edge e of T, and let T_1, T_2 be the two rooted subtrees of T, whose roots are adjacent to the root vertex on e. Let π_1, π_2 be the marginal distributions of π restricted to the leaves in T_1 and T_2 respectively.



In view of Lemma 1 we have:

$$F_X(T, \pi, x) = \sum_{A \cap B = X} F_A(T_1, \pi^1, x) F_B(T_2, \pi^2, x) + \sum_{\substack{A \cap B = \emptyset \\ A \cup B = X}} x F_A(T_1, \pi^1, x) F_B(T_2, \pi^2, x), \qquad (19)$$

since we add 0 or 1 to the sum of the lengths of T_1 and T_2 depending on whether $A \cap B \neq \emptyset$ or $A \cap B = \emptyset$ by the forward recursion version of Fitch's algorithm. The summation is over pairs A, B satisfying the stated conditions. Note that if T_1 has just one leaf i then $F_A(T_1, \pi^1, x) = \pi_i^\alpha$, if $A = \{\alpha\}$, while $F_A(T_1, \pi^1, x) = 0$, if $|A| > 1$.

Thus, one can efficiently calculate the polynomials $\{F_X(T,\boldsymbol{\pi},x) : X \neq \emptyset\}$ by starting from the leaves, and working up to the root and storing all the intermediate polynomials generated in the construction. Then $\mathbb{P}[T,\boldsymbol{\pi},k]$ is simply the coefficient of x^k in $\sum_{X\neq\emptyset} F_X(T,\boldsymbol{\pi},x)$.

Next we consider the complexity of this algorithm. For the subtree T_v below v we must compute equation (19) for each $X \neq \emptyset$, *i.e.* $2^c - 1$ times. The terms $F_A(T_1,\boldsymbol{\pi}^1,x)$ and $F_B(T_2,\boldsymbol{\pi}^2,x)$ can be multiplied in time proportional to $(\max L(T_1))(\max L(T_2))$ steps which is less than or equal to the product of the number of leaves of the two trees. This product associated with T_v will be denoted by $n(v)$. The number of solutions to $A * B = X$ is

$$2^{|X|} + \sum_{i\geq 0} \binom{c-|X|}{i} 2^{c-|X|-i} = 2^{|X|} + 3^{c-|X|} \leq 3^c.$$

Thus $F_X(T,\boldsymbol{\pi},x)$ can be computed from $\{F_A(T_1,\boldsymbol{\pi}^1,x), A \neq \emptyset\}$ and $\{F_B(T_2,\boldsymbol{\pi}^2,x), B \neq \emptyset\}$ in $O((2^{|X|}+3^{c-|X|})n(v))$ steps. Noting that $\sum_X 2^{|X|}+3^{c-|X|} = 3^c+4^c$, we see that $\{F_X(T_v,\boldsymbol{\pi},x), X \neq \emptyset\}$ can be obtained from the previous sets in $O(4^c n(v))$ steps. A straightforward inductive argument shows that for any tree T with n leaves

$$\sum_{\substack{v\in V[T]\\ \deg v>1}} n(v) \leq \binom{n}{2}.$$

(This upper bound being realized by a caterpillar tree). Thus $\{F_X(T_v,\boldsymbol{\pi},x), X \neq \emptyset\}$ can be computed in $O(4^c n^2)$ steps.

This recursive description allows the distribution of $L(T)$ to be effectively calculated, even when n is quite large (say 10^3). At each vertex we must compute F_X for all non-empty subsets X. For c colors this means $2^c - 1$ values of X. While this is not a problem for $c = 4$ it would be for $c = 20$. In any case, this recursion appears to be of little help in determining what the limiting distribution is as $n \to \infty$.

Now we turn to a special algorithm designed to directly compute $\mathbb{E}[L] = \mathbb{E}[L(T)]$ and $V[L] = V[L(T)]$. First we compute $\mathbb{E}[L]$.

Subdivide an edge of T, root T at this newly created vertex, and direct all the edges of T away from this root. For any nonleaf vertex v of T let T_v denote the subtree of T consisting of those vertices which are descendents of v. Note that T_v is a rooted binary tree, with v as its root. Let

$$P_X[T_v] = \mathbb{P}[S(T) = X].$$

Since v is not a leaf it has two immediate descendent vertices v' and v'' and then by Lemma 1,

$$P_X[T_v] = \sum_{A*B=X} P_A(T_{v'})P_B(T_{v''}),$$

where $A * B = A \cap B$ if $A \cap B \neq \emptyset$ and $A \cup B$ if $A \cap B = \emptyset$. Thus we can calculate the set $\Omega = \{P_X(T_v) : X \neq \emptyset, \deg(v) > 1\}$ starting with the initial conditions on the leaves:

$$P_X[i] = \begin{cases} \pi_i^\alpha, & \text{if } X = \{\alpha\}, \\ 0, & \text{if } |X| > 1. \end{cases}$$

Constructing Ω requires $O(n4^c)$ steps, as T has $O(n)$ vertices. Now, by Lemma 1,

$$\mathbb{E}[L] = \sum_{\deg(v)>1} \sum_{A\cap B=\emptyset} P_A(T_{v'})P_B(T_{v''})$$

so that $\mathbb{E}[L]$ can be calculated in a further $O(n)$ steps from the set Ω constructed above so $\mathbb{E}[L]$ is computable in $O(n4^c)$ steps.

Recall that

$$L = L' + L'' + D,$$

where

$$D = \begin{cases} 1, & \text{if } S(T') \cap S(T'') = \emptyset, \\ 0, & \text{otherwise.} \end{cases}$$

Now L' and L'' are independent and

$$V[L] = V[L'] + V[L''] + 2Cov(D, L' + L'') + V[D] \tag{20}$$

The quantity $V[D]$ is handled next. Since D has values 0 and 1, $V[D] = \mathbb{P}[D = 1]\{1 - \mathbb{P}[D = 1]\}$ and

$$\mathbb{P}[D = 1] = \sum_{A\cap B=\emptyset} \mathbb{P}[S(T') = A]\mathbb{P}[S(T'') = B].$$

The probabilities are computed in the algorithm for $\mathbb{E}[T]$. $V[D]$ is now computable in $O(3^c)$ additional steps since

$$\sum_i \binom{c}{i} 2^{c-i} = 3^c.$$

Now we consider the covariance term in equation (20).

$$Cov(D, L' + L'') = \mathbb{E}[DL'] + \mathbb{E}[DL''] - \mathbb{E}[D](\mathbb{E}[L'] + \mathbb{E}[L'']).$$

The terms not already considered are $\mathbb{E}[DL']$ and $\mathbb{E}[DL'']$.

$$\begin{aligned}
\mathbb{E}[DL'] &= \mathbb{E}[L'|D=1]\mathbb{P}[D=1] \\
&= \sum_k k\mathbb{P}[L' = k \cap D = 1] \\
&= \sum_{A \cap B = \emptyset} \sum_k k\mathbb{P}[L' = k \cap \{S(T') = A\} \cap \{S(T'') = B\}] \\
&= \sum_{A \cap B = \emptyset} \mathbb{P}[S(T'') = B] \sum_k k\mathbb{P}[L' = k \cap \{S(T') = A\}] \\
&= \sum_{A \cap B = \emptyset} \mathbb{P}[S(T'') = B] F(T', A),
\end{aligned}$$

where $F(T', A)$ is defined by the last equation.

$$\begin{aligned}
F(T, A) &= \sum_k k\mathbb{P}[\{L = k\} \cap \{S(T) = A\}] \\
&= \sum_{B * C = A} \sum_{k_1, k_2} (k_1 + k_2 + \mathbb{I}(B \cap C \neq \emptyset))\mathbb{P}[\{L_1 = k_1\} \cap \{S(T') = B\} \cap \{L_2 = k_2\} \cap \{S(T'') = C\} \\
&= \sum_{B * C = A} \sum_{k_1, k_2} (k_1 + k_2 + \mathbb{I}(B \cap C \neq \emptyset))\mathbb{P}[\{L_1 = k_1\} \cap \{S(T') = B\}]\mathbb{P}[\{L_2 = k_2\} \cap \{S(T'') = C\}
\end{aligned}$$

Breaking the sum into 3 parts we obtain

$$F(T, A) = \sum_{B * C = A} \{F(T', B)\mathbb{P}[S(T'') = C] + F(T'', C)\mathbb{P}[S(T') = B] + \mathbb{I}(B \cap C \neq \emptyset)\mathbb{P}[S(T') = B]\mathbb{P}[S(T'') = C$$

Clearly this last equation allows us to recursively compute $\mathbb{E}[DL']$, and similarly $\mathbb{E}[DL'']$. Hence $Cov(D, L' + L'')$ and hence equation (20) for $V[L]$ can be computed. The number of pairs B, C such that $B * C = A$ is bounded by $2^{|A|} + 3^{c-|A|}$, and so $V[L]$ can be computed in time $O(4^c n)$.

5 Example

Consider a "bush" of height h - that is, any binary tree T_h for which the path from any leaf to the root contains exactly h edges. It is easily seen that a bush of height h has 2^h leaves. Now suppose the leaves of T are independently and identically bicolored from $\{\alpha, \beta\}$ with $p \geq 0.5$ the probability that each leaf is colored α. If $p = 0.5$, then, from Charleston and Steel (1994),

$$\mathbb{P}[S(T_h) = \{\alpha\}] = \mathbb{P}[S(T_h) = \{\beta\}] = \frac{1}{3}\left(1 - (-0.5)^h\right)$$

and so

$$\lim_{h\to\infty} \mathbb{P}[S(T_h) = X] = \frac{1}{3} \text{ for } X = \{\alpha\}, \{\beta\}, \{\alpha, \beta\}.$$

Now suppose $p > 0.5$. In this case, we will show that

$$\lim_{h\to\infty} \mathbb{P}[S(T_h) = X] \text{ is 1 if } X = \{\alpha\}, \text{ and is 0 otherwise!}$$

Thus, the distribution of $S(T_h)$ can be asymptotically degenerate even when the leaf coloration distributions are *i.i.d.* and bounded away from degenerate (*i.e.* satisfy (10)).

In order to substantiate our claim, write $\mathbb{P}[S(T_h) = X]$ as $p_X(h)$. Deleting from T_h the root and its two incident edges gives two bushes of height $h-1$, thereby providing the system of simultaneous recursions:

$$\begin{aligned}
p_{\{\alpha\}}(h) &= p^2_{\{\alpha\}}(h-1) + 2p_{\{\alpha\}}(h-1)p_{\{\alpha,\beta\}}(h-1) \\
p_{\{\beta\}}(h) &= p^2_{\{\beta\}}(h-1) + 2p_{\{\beta\}}(h-1)p_{\{\alpha,\beta\}}(h-1) \\
p_{\{\alpha,\beta\}}(h) &= p^2_{\{\alpha,\beta\}}(h-1) + 2p_{\{\alpha\}}(h-1)p_{\{\beta\}}(h-1)
\end{aligned}$$

Letting $p_X = \lim_{h\to\infty} p_X(h)$, we have:

$$p_{\{\alpha\}} = p^2_{\{\alpha\}} + 2p_{\{\alpha\}}p_{\{\alpha,\beta\}}$$

$$p_{\{\beta\}} = p_{\{\beta\}}^2 + 2p_{\{\beta\}}p_{\{\alpha,\beta\}}$$

$$p_{\{\alpha,\beta\}} = p_{\{\alpha,\beta\}}^2 + 2p_{\{\alpha\}}p_{\{\beta\}}$$

which, it can readily be checked, has only four solutions $(p_{\{\alpha\}}, p_{\{\beta\}}, p_{\{\alpha,\beta\}})$ for which the components sum to 1, namely (1,0,0), (0,1,0), (0,0,1) and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Now, consider $D(h) = p_{\{\alpha\}}(h) - p_{\{\beta\}}(h)$. We have:

$$D(h) = D(h-1)[p_{\{\alpha\}}(h-1) + p_{\{\beta\}}(h-1) + 2p_{\{\alpha,\beta\}}(h-1)]$$

But now, $p_{\{\alpha\}}(h-1) + p_{\{\beta\}}(h-1) = 1 - p_{\{\alpha,\beta\}}(h-1)$, and so,

$$D(h) = D(h-1)[1 + p_{\{\alpha,\beta\}}(h-1)].$$

Thus, for all h, $D(h) > D(1)$, and $D(1) > 0$, since $p > 0.5$, so the only possible value for $(p_{\{\alpha\}}, p_{\{\beta\}}, p_{\{\alpha,\beta\}})$ is (1,0,0), as claimed.

6 Discussion

Theorems 1 and 3 apply to one position of n aligned homologous sequences. The common assumption in phylogenetic analysis is that positions are i.i.d. For aligned sequences the critera is the sum of parsimony scores over all positions. In this case the central limit theorem for i.i.d. random variables applies. Our Theorem 3 shows that the individual terms making up this sum are themselves approximately normal, hence an excellent fit of this sum to the normal is expected. In contrast, Theorem 2 gives large deviation bounds for the tree that attains the minimum length (maximum parsimony) score over k positions of n aligned homologous sequences where no such central limit theorem applies.

We note finally that the central limit theorem is not true in general for non-binary trees. For a simple counterexample, take the star tree; that is, the tree with $n+1$ vertices and n edges all of the form $\{v_0, v\}$ for a distinguished (center) vertex v_0. For the star tree, the minimum coloration extending a given leaf coloration is the one in which the center vertex is colored the most frequent

color; the length L is therefore the number of leaves not colored with the most frequent color. Hence, in the case of two colors α and β, for $k \leq n/2$, $L = k$ when there are k or $n-k$ leaves with color α. Assigining each color with equal probability at each leaf, $P[L = k] = \binom{n}{k} 2^{1-n}$ if $0 \leq k \leq n/2$ and 0 otherwise. As this distribution drops from its most probable value to 0 when increasing $[n/2]$ by 1, it cannot converge to the normal.

Acknowledgement. We thank Michael J. Steele for describing the Martingale-style approach for bounding variances from below (used in the proof of Proposition 1).

References

Alon, N. and Spencer, J.H. *The probabilistic method*, John Wiley and Sons, New York 1992.

Archie, J. and Felsenstein J. The number of evolutionary steps on random and minimum length trees for random evolutionary data, *Theor. Pop. Biol.* **43**, 52-79, 1993.

Carter, M., Hendy, M., Penny, D., Szekely, L.A. and Wormald, N.C. On the distribution of lengths of evolutionary trees, *SIAM J. Disc. Math.* **3**(1), 38-47, 1990.

Charleston, M. and Steel, M.A. Five surprising properties of parsimoniously colored trees, *Bull. Math. Biol.*, 1994 (in press).

Durrett, R. *Probability: Theory and Examples*, Wadsworth & Brooks/Cole, Belmont, CA. 1991.

Fitch, W. M. Toward defining the course of evolution: minimum change for a specific tree topology, *Syst. Zool.* **20**, 406-416, 1971.

Hartigan, J.A. Minimum mutation fits to a given tree, *Biometrics* **29**, 53-65, 1973.

Moon, J.W. and Steel, M.A. A limiting distribution for parsimoniously bicolored trees, Appl. Math. Lett. **6**(4), 5-8, 1993.

Serfling, R.J. *Approximation Theorems of Mathematical Statistics*, John Wiley and Sons, New York 1980.

Steel, M.A. Decompositions of leaf-coloured binary trees, *Adv. Appl. Math.* **14**(1), 1-24, 1993.

Steel, M.A. Distributions on bicoloured binary trees arising from the principle of parsimony, *Discr. Appl. Math.* **41**, 245-261, 1993.

Steele, J.M. An Efron-Stein inequality for nonsymmetric statistics, *Ann. Stat.* **14**, 2,753-758, 1986.